Anova (Analysis of Variance)

- M1 : $Y_i = \beta_0 + e_i \xrightarrow{RSS} \Sigma(Y_i - \bar{Y})^2 = SYY$
- M2 : $Y_i = \beta_0 + \beta_1 X_i + e_i \xrightarrow{RSS} SYY - \frac{SXY^2}{SXX}$

$$SSreg = RSS_{M1} - RSS_{M2} = \frac{SXY^2}{SXX} = \begin{cases} \text{large} \rightarrow \text{favors M2} \\ \text{small} \rightarrow \text{favors M1} \end{cases}$$

$$\boxed{\begin{array}{l} H_0: \text{M1 is the right model} \\ H_A: \text{M2 is the right model} \end{array}}$$

$$SSreg = \left(\frac{\Sigma(X_i - \bar{X}) Y_i}{\sqrt{SXX}} \right)^2 \overset{①②}{\sim} \underline{\sigma^2} \chi_1^2$$

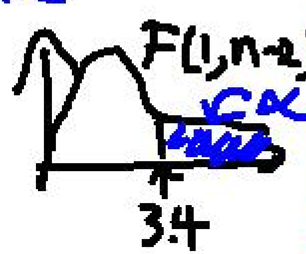
↑ unknown

$$\text{F-stat} = \frac{SSreg}{\hat{\sigma}^2} \overset{③}{\sim} F(1, n-2)$$

$SSreg \leftarrow \cancel{\sigma^2} \chi_1^2$

$\frac{RSS}{n-2} \rightarrow \hat{\sigma}^2 \leftarrow \frac{\Sigma_1^n \hat{e}_i^2}{n-2} \sim \cancel{\sigma^2} \chi^2_{n-2} / n-2$

eg data $\rightarrow$ F-stat $= 3.4 \xrightarrow[R]{\text{use}}$



If $\alpha < 0.05 \Rightarrow$ reject H_0

$\alpha \geq 0.05 \Rightarrow$ not reject H_0

Stat facts

① $\sum a_i Y_i \xrightarrow{CLT}$ Normal

② $X_i \sim N(0,1)$

$\Rightarrow X_1^2 + X_2^2 + \cdots + X_k^2 \sim \chi_k^2$

③ $X_1 \sim \chi_m^2$

$X_2 \sim \chi_n^2$

$\Rightarrow \frac{X_1/m}{X_2/n} \sim F(m,n)$

④ $X_1 \sim N(0,1)$

$X_2 \sim \chi_m^2$

$\Rightarrow X_1 / \sqrt{X_2/m} \sim t_m$

2.6 Comparing models: Analysis of variance (ANOVA)

- ANOVA table: a break-down of squares (variation)

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg\,/\,1$	$MSreg\,/\,\hat{\sigma}^2$	
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS\,/\,(n-2)$		
Total	$n-1$	SYY			

$SYY - SXY^2/SXX$

SYY = RSS + SSreg $\left(\frac{SXY^2}{SXX}\right)$

$$\sum_{i=1}^{n}[y_i - \bar{y}\,]^2 = \boxed{\sum_{i=1}^{n}[y_i - \hat{y}_i\,]^2} + \boxed{\sum_{i=1}^{n}[\hat{y}_i - \bar{y}\,]^2}$$

$$TSS = SYY = RSS + SSreg$$

ch 8

fitted data

Variation of the data	Variation not explained by regression	Variation explained by regression

2.6 Comparing models: Analysis of variance (ANOVA)

- ANOVA table: a break-down of squares (variation)

Source	df	SS	MS	F	p-value
Regression	1	$SSreg$	$SSreg/1$	$MSreg/\hat{\sigma}^2$	
Residual	$n-2$	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	SYY			

$$\sum_{i=1}^{n}[y_i - \bar{y}]^2 = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 + \sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2$$

$$TSS = SYY = RSS + SSreg$$

TSS = SYY	RSS	SSreg
Variation of the data	**Variation not explained by regression**	**Variation explained by regression**

$$R^2 = \frac{SSreg}{SYY} \in [0, 1] = \begin{cases} \approx 1 & \text{Reg explained a lot} \\ \approx 0 & \text{Lots of error} \end{cases}$$

2.7 Coefficient of Determination, R^2

- Definition

$$R^2 = \frac{SSreg}{SYY} \leftarrow \frac{SXY^2}{SXX}$$

  - Proportion of variability explained by regression
- Scale-free one number summary of strength of relationship between x and y.
- Connections to sample correlation r^2_{xy}

$$R^2 = \frac{SSreg}{SYY} = \frac{(SXY)^2}{SXX\ SYY} = r^2_{xy}$$

$$\left[\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\,\Sigma(y_i-\bar{y})^2}}\right]^2$$

- R^2 is always between 0 and 1.
  - Close to 1→ good fit
  - Close to 0→ bad fit

$$\boxed{n = 12}$$

$$F = 2.4 = \frac{SSreg}{RSS/10}$$

$$\text{Want } R^2 = \frac{SSreg}{SYY}$$

$$\boxed{SYY = SSreg + RSS}$$

$$= \frac{SSreg}{SSreg + RSS}$$

$$= \frac{SSreg/RSS}{SSreg/RSS + 1}$$

$$= \frac{0.24}{0.24 + 1} = \frac{24}{124} = \frac{6}{31}$$

2.8 Confidence intervals and tests

- slope: β_1
- Testing:

$$\begin{aligned} &\text{NH:} \quad \beta_1 = \beta_1^*, \quad \beta_0 \text{ arbitrary} \\ &\text{AH:} \quad \beta_1 \neq \beta_1^*, \quad \beta_0 \text{ arbitrary} \end{aligned}$$

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\boxed{\mathrm{se}(\hat{\beta}_1)}} \qquad t\text{-distribution with } n-2 \text{ df}$$

← df n2

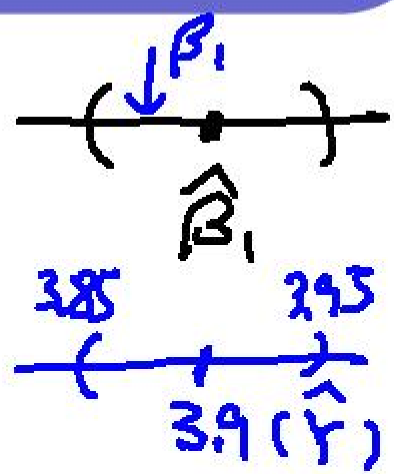
where $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SXX}}$

- (1-α)x100% confidence interval:

$$\hat{\beta}_1 - t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1)$$

$Y_i = \beta_0 + \beta_1 X_i + e_i$ ← individual variation

2.8 Confidence intervals and tests



- Regression model:
  - $E(Y|X=x)= \beta_0 + \beta_1 x$
- Quantities of interests
  - Intercept: β_0
  - Slope: β_1
  - individual → Prediction: If we observe x_*, what is the y?
  - average → Fitted value: $E(Y|X=x)$ for different values of x
- Confidence intervals give estimates for the above quantities of interests

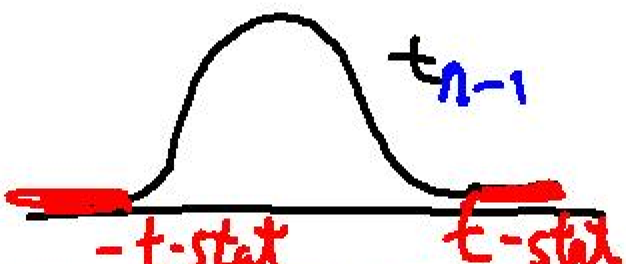
Remember STAT2001 $(X_1 \ldots X_n) \sim$ true mean μ

$$\boxed{\begin{array}{l} H_0 : \mu = 3 \\ H_A : \mu \neq 3 \end{array}}$$

Testing

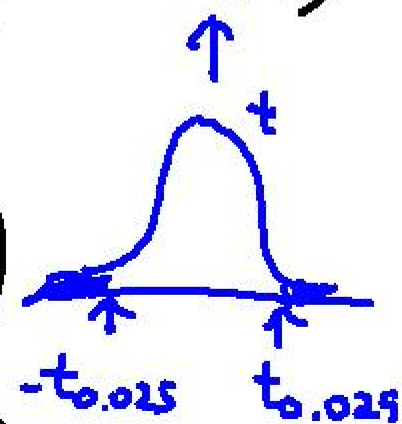
$$t = \frac{\bar{x} - 3}{\sqrt{Var(\bar{x} - 3)}} = \frac{\bar{x} - 3}{\hat{\sigma}/\sqrt{n}}$$

$$\sim \frac{\text{Normal}}{\sqrt{\chi^2}} \sim t\text{-dist}$$



C.I. for μ

$$0.05 = P\left(-t_{0.025} < \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} < t_{0.025}\right)$$



$$\Rightarrow 0.05 = P\left(\bar{x} - t\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + t\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

$$\Rightarrow \mu \in \underbrace{\left(\bar{x} - t\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + t\frac{\hat{\sigma}}{\sqrt{n}}\right)}_{C.I.}$$

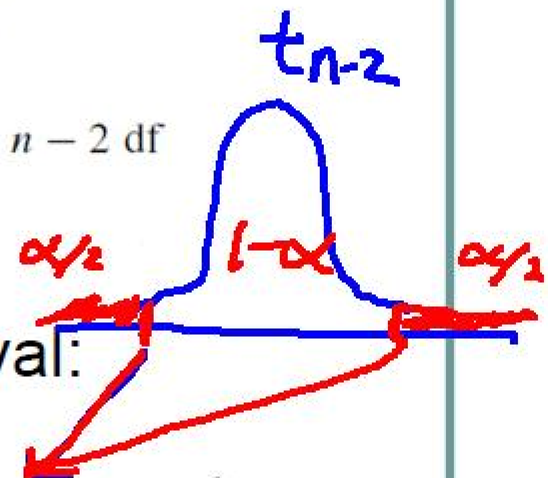
2.8 Confidence intervals and tests

- slope: β_1
- Testing:

$$\text{NH:} \quad \beta_1 = \beta_1^*, \quad \beta_0 \text{ arbitrary}$$
$$\text{AH:} \quad \beta_1 \neq \beta_1^*, \quad \beta_0 \text{ arbitrary}$$

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\text{se}(\hat{\beta}_1)} \qquad t\text{-distribution with } n-2 \text{ df}$$

where $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SXX}}$



- (1-α)x100% confidence interval:

$$\hat{\beta}_1 - t(\alpha/2, n-2)\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t(\alpha/2, n-2)\text{se}(\hat{\beta}_1)$$

$$\mathrm{Var}\left((\beta_0-\hat{\beta}_0) + (\beta_1-\hat{\beta}_1)X_* + e \right)$$

$$= \mathrm{Var}(\beta_0-\hat{\beta}_0) + X_*^2\,\mathrm{Var}(\beta_1-\hat{\beta}_1) + \sigma^2$$

$$+\, 2\,\mathrm{Cov}(\beta_0-\hat{\beta}_0,\ \beta_1-\hat{\beta}_1)$$

$$+\, \cancel{2\,\mathrm{Cov}(\beta_0-\hat{\beta}_0,\ e)}$$

$$+\, \cancel{2\,\mathrm{Cov}(\beta_1-\hat{\beta}_1,\ e)}$$

indivial

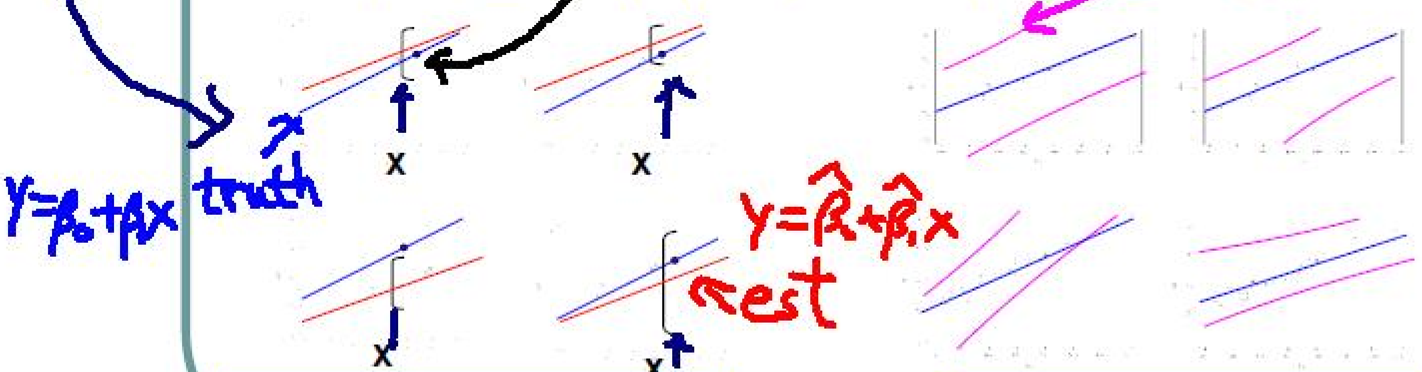
from new obs

from past data

2.8 Confidence intervals and bands

- Confidence interval (at each point x)
  - For each of x, P(E(Y|X=x) in C.I.)=1-α
- Confidence band (for the entire line)
  - P(For all x, E(Y|X=x) in C.B.)= 1-α

C.B



For n C.I.s, n(1-α) of them covers the true value at x

For n C.B.s, n(1-α) of them covers the whole true regression line